DOCUMENT RESUME

ED 127 529 CG 010 757

AUTHOR Frary, Robert B.; Lowry, Stephen R.
TITLE Misinformation, Reliability and Item Discrimination Indices on Multiple Choice Tests.
PUB DATE Apr 76
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, California, April 19-23, 1976); not available in hard copy due to marginal legibility of original document

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *Bias; *Correlation; Educational Research; *Multiple Choice Tests; Research Projects; Speeches; *Test Construction; *Test Reliability; Tests; *Theories

ABSTRACT

This paper presents theory concerning the relationships between reliability, misinformation and item discrimination coefficients. It is shown that, to the extent that misinformation rather than ignorance causes examinees to miss multiple-choice items, higher item discrimination coefficients and lower difficulty indices may be expected. Data were collected which partially confirmed the prevalence of these outcomes in typical college classroom testing situations involving six tests and 210 examinees. The implications of the findings are discussed with respect to commonly used test construction procedures. Specifically, a caution is voiced concerning possible biasing of tests to penalize misinformation more than ignorance when this approach is inappropriate. (Author)

MISINFORMATION, RELIABILITY AND ITEM DISCRIMINATION INDICES ON MULTIPLE-CHOICE TESTS

Robert B. Frary and Stephen R. Lowry
Virginia Polytechnic Institute
and State University

Paper presented at the Annual Meeting of
the American Educational Research Association
San Francisco, California, April, 1976
(Session 12.14: Small Roundtable, Division D)

MISINFORMATION, RELIABILITY AND ITEM DISCRIMINATION INDICES ON MULTIPLE-CHOICE TESTS

ABSTRACT

Robert B. Frary and Stephen R. Lowry
Virginia Polytechnic Institute
and State University

This paper presents theory concerning the relationships between reliability, misinformation and item discrimination coefficients. It is shown that, to the extent that misinformation rather than ignorance causes examinees to miss multiple-choice items, higher item discrimination coefficients and lower difficulty indices may be expected. Data were collected which partially confirmed the prevalence of these outcomes in typical college classroom testing situations involving six tests and 210 examinees. The implications of the findings are discussed with respect to commonly used test construction procedures. Specifically, a caution is voiced concerning possible biasing of tests to penalize misinformation moreso than ignorance when this approach is inappropriate.

Misinformation, Reliability and Item Discrimination Indices on Multiple-Choice Tests

Robert B. Frary and Stephen R. Lowry
Virginia Polytechnic Institute and State University

The concepts of misinformation and ignorance are critical for interpreting the results of testing. In academic testing, for example, prevalence of low scores due to misinformation suggests an entirely different approach to remediation than low scores resulting from ignorance. In professional licensing examinations, misinformation probably represents much stronger grounds for denial of licensing than does ignorance. An active medical practitioner may seek consultation with colleagues when in doubt regarding treating a patient. In contrast, a misinformed practitioner may make a fatal mistake.

On a multiple-choice test question, it is plausible to define ignorance and misinformation according to an examinee's strategy or behavior in answering the question. When the examinee does not know the answer with a substantial degree of certainty, yet intends to answer it nevertheless, the following behavior is hypothesized. First he eliminates choices he believes to be wrong with a substantial degree of certainty. Then he guesses among the remaining choices. If the eventual choice is incorrect, this outcome may be categorized as being due either to ignorance or to misinformation on the following basis:

Ignorance. The right answer was one of the choices among which the examinee guessed.

Misinformation. The right answer was one of the choices eliminated with a substantial degree of certainty.

It will be shown that item selection procedures typically employed in test

development tend to bias tests so that low scorers are likely to display misinformation moreso than ignorance.

Item Difficulty

A typical problem in test development is finding good, yet difficult, items. In order to maximize internal consistency for a test designed to measure a wide range of performance, a substantial proportion of difficult items is required, that is, items answered correctly by less than half of the examinees. Items of this sort are likely to reflect misinformation moreso than ignorance. This statement follows because, in the absence of misinformation and with a small proportion of examinees knowing the answer with assurance, guessing success may make the item appear to be of only medium difficulty. The following hypothetical response proportions illustrate this point for four items, each of which is known with assurance by only one-tenth of the examinees:

	Test Item			
	1	2	3	4
A. Proportion knowing answer with assurance	.1	.1	.1	.1
B. Proportion with misinformation	.0	.0	.4	.4
C. Proportion of ignorant examinees	.9	.9	.5	.5
D. Average probability of correct guess for ignorant examinees	.2	.4	.2	.4
E. Proportion of correct answers from guessing (C x D)	.18	.36	.1	.2
F. Total proportion correct (A + E)	.28	.46	.20	.30

In seeking difficult items, there would be a substantial tendency to choose items like 3, which have higher proportions wrong due to misinformation. However, items like 1 and 3, on which guessing would be essentially random (assuming five choices), are probably quite rare. Difficult items are more likely to be selected from ones like 2 and 4, on which probability of a correct guess is higher. Again the items representing more misinformation seem more likely to be chosen on the basis of difficulty.

Item Correlation with Total and Criterion Scores

Other than difficulty level, the main basis for selecting one item over another is correlation with total score or a criterion. Again, items representing misinformation are more likely to appear superior in this respect. To understand why this result follows, consider two items, both of which can be answered with assurance by high scorers on the test itself or the criterion. In contrast, low scorers on the test or criterion are ignorant on one item and misinformed on the other. The item representing ignorance will be answered correctly by a proportion of the ignorant low scorers, thus decreasing its correlation with total score or the criterion in comparison with the other item, which is answered incorrectly by all of these same low scorers.

Empirical Investigation

Determination of the validity of the theory developed above requires knowledge for each item of a test in the preliminary stages of development of the proportion of examinees with misinformation. Then across test items under consideration for retention this proportion should correlate with item selection statistics.

Data for the study came from six college-level biology tests consisting of 20 to 35 4-choice items. The tests were administered to three groups (two

tests per group) of approximately 40 to 80 students each. Students were instructed to respond according to the mode suggested by Coombs and others (1956) in which examinees mark the choices they believe incorrect with a strong penalty for inadvertently marking the correct answer. Course grades were assigned on the basis of this scoring procedure, so that students should have been strongly motivated to avoid marking choices about which they were unsure. (Examinees received one point for each correctly identified wrong answer and a three point penalty for marking the right choice.) In addition, examinees recorded their order of elimination of wrong choices and their best guesses as to the correct answers to produce the usual number-right scores.

Whenever an examinee inadvertently marked a correct choice as incorrect, it was assumed that he had misinformation to some degree. Further, it was assumed the earlier he marked the correct choice, the greater the degree of his misinformation, that is, the more confident he was that the right choice was wrong. However, an artifact in the response procedure tended to bias this measure. Examinees apparently tended to eliminate an earlier appearing choice sooner than a later one when there was an approximately equal degree of confidence that they both were wrong. This phenomenon was detected by noting a substantial correlation across items between choice position for wrong choices and mean order of elimination on all six tests. To correct for this bias, the mean order of elimination of right choices was standardized as the deviation of this mean for each item from the same mean over all items with the same choice position. For a procedural explanation of this standarization, consider a hypothetical item for which the correct answer is the second choice. Perhaps ten examinees inadvertently eliminated this choice and their average order of elimination of

this choice is 2.1. For all items for which the second choice is the correct answer, the mean order of inadvertent eliminations is then calculated to be 2.5. The deviation, .4, is then used to express the degree or intensity of misinformation associated with the item in question. Its positive value suggests that examinees were more prone than on other items to eliminate the right choice.

The above procedures made possible computation of three measures of misinformation for each test item:

ANIS: Proportion of examinees with misinformation, an absolute measure of misinformation.

RMIS: Proportion of examinees missing the item who displayed misinformation, a relative measure of misinformation.

IMIS: Deviation of mean order of elimination of right answer from mean order of elimination of right answer across all items with same choice position for right answer, a measure of intensity of misinformation.

For each item of the six tests, the following item quality indices were computed:

DISC: Item discrimination coefficient (point biserial).

DDEV: Deviation of item difficulty index from .5.

QUAL: Overall item quality index

$$\sqrt{4(DDEV)^2 + (DISC - 1)^2}$$

as recommended by Davis (1964).

Initial inspection of statistics from the six tests revealed a number of exceedingly easy items. The presence of a number of such items would have resulted in a very skewed distribution of the number of examinees with misinformation (across items), since it is not possible to have a large proportion misinformed when very few miss the item. Also such items are

easily avoided in test development. Therefore, items answered correctly by more than 90 percent of the examinees were dropped from the study. The tests were rescored, and all statistics reported reflect only the residual items. These remaining items yielded test statistics as shown in Table 1. As the reader may judge, these statistics appear typical of those encountered in the early stages of test development. However, one statistic usually not available to test developers, correlation (across examinees) between number of items missed due to misinformation and raw score was surprising. For each of the six tests, this correlation, though negative, was of only moderate size, suggesting that misinformation did not reside so extensively among low scorers as assumed in the discussion concerning discrimination coefficients. Further computation revealed that the correlations between proportion of items missed by each examinee due to misinformation and total score were near zero. Therefore, for the six tests under consideration, data analysis then focused on the extent to which item selection criteria might be related to misinformation in the absence of satisfying the assumption that extent of misinformation was strongly related to total score.

For each test intercorrelations among the three quality and three misinformation indices were computed (across items) as shown in Table 2. Correlations between AMIS (absolute proportion of misinformation) and DISC (discrimination coefficients) are generally negative, that is, low discrimination coefficients (poor items) seem to be associated with higher numbers of misinformed examinees. This result is contrary to what was expected and probably reflects the fact that misinformation was not substantially concentrated among lower scorers. In contrast correlations between AMIS and DDEV (absolute

value of deviation of item difficulty index from .5) are substantially negative as suggested by earlier discussion. This outcome reflects the fact that a majority of items had difficulty indices above .5. As a result, those with high AMIS values usually had difficulty indices closer (yet still above) .5. This situation is partly artifactual in that tests with more difficult items might not have yielded this relationship. Yet it is realistic. Development of tests or item banks typically involves greater elimination of easy items than difficult, especially if means of the neighborhood of 50 to 60 percent are desired. The total effect on QUAL (Davis item quality index) of item difficulty and discrimination items with difficulty indices substantially above or below .5 tend to have high (poor) quality indices even when they have good discrimination coefficients. Accordingly the correlations between AMIS and QUAL are generally negative in spite of the negative correlations between AMIS and DISC.

Correlations between RMIS and IMIS and the three quality indices reveal no obvious patterns of statistical significance. A correlation of .44 ($p < .1$) for Test 2 between IMIS and DISC does suggest that items on which the misinformation is more intense have higher discrimination indices. Also, for Test 2, a correlation of .47 ($p < .05$) between RMIS and DDEV is contrary to what might have been expected, suggesting that items with difficulties deviating moreso from .5 have higher relative misinformation measures. However, there is no significance over the 109 items of the study for any correlation involving RMIS and IMIS.

Discussion

The data of the study support the theoretical conclusions only in one respect, namely, that items with difficulty indices nearer .5 tend to represent higher

than average proportions of misinformation across examinees. Of course this outcome is at least partly due to a preponderance of items with difficulty indices above .5 among those available for analysis. Misinformation naturally reduces the number of correct responses, making an otherwise easy item look better statistically.

Nevertheless, the implications for test development are not trivial. In a practical sense, there is every reason to believe that items representing misinformation moreso than ignorance have a greater chance of selection from many item pools. Of course, the substantive nature of the items needs to be considered. For example, some tests, usually requiring numerical computation of answers, may have all wrong choices representing incorrect solutions arising from specific mistakes (misinformation), virtually requiring the ignorant examinee to guess at random or omit the item.

In spite of and because of the somewhat inconclusive nature of what has been precented, there are substantial implications for further research. In addition to knowledge scores from tests, examinees might well earn misinformation and ignorance scores or perhaps a single additional score representing the ratio of misinformation to ignorance. Validity studies might then reveal the meaning of these scores with respect to a variety of criteria.

Other research might investigate various approaches to measuring misinformation and ignorance. One approach might involve use of the "I don't know" responses used for various standardized tests. The proportion of items missed from among those attempted might be related to misinformation and the number of "I don't know" responses to ignorance. Various refinements to the methods presented in this paper are also possible. For example, multiple forms of the tests might be used with random ordering of choices within an item to avoid the problem of bias due to earlier misinformed elimination of right choices which appeared earlier in the list of choices for an item.

REFERENCES

Coombs, C.H., Milholland, J.E., & Womer, F.B. The assessment of partial knowledge. *Educational and Psychological Measurement*, 1956, *16*, 13-37.

Davis, F.B. *Educational measurements and their interpretation*, Belmont, California: Wadsworth, 1964.

Table 1

Statistics for Six Biology Tests

	1	2	3	4	5	6
Number of Examinees	74	74	42	42	78	80
Number of Items[1]	19	20	18	22	22	14
Mean Item Difficulty	.71	.75	.74	.59	.52	.68
Mean Item Discrimination	.26	.36	.32	.41	.38	.36
KR-20	.58	.82	.47	.67	.69	.47
Mean Order of Elimination for Misinformed Responses[2]	2.0	1.9	1.9	1.9	2.1	2.1
Mean Proportion of Examinees per Item with Misinformation	.12	.09	.13	.20	.25	.25
Mean Proportion Per Item of Wrong Choices Due to Misinformation	.38	.41	.48	.50	.56	.79
Mean Absolute Deviation of Item Difficulty Indices from .5	.22	.26	.24	.16	.16	.22
Mean Item Quality Index	.80	.78	.78	.77	.72	.79
Correlation Between Raw Score and Number of Items Missed Due to Misinformation	-.41	-.47	-.37	-.45	-.46	-.47
Correlation Between Raw Score and Proportion of Items Missed Due to Misinformation	-.07	-.13	-.17	-.05	.03	-.01

[1]Does not include items with difficulty indices above .9, which were not used in the study (see text).

[2]All items had four choices used with approximately equal frequency across items. Hence misinformed eliminations tended to occur sooner than the choice position of the correct answer.

Table 2

Correlations Between Item Quality
and Misinformation Indices

	TEST 1	2	3	4	5	6	ALL ITEMS COMBINED
Number of Items:	19	20	18	22	22	14	109
$r_{AMIS, DISC}$	-.16	.08	-.45*	-.38*	-.08	-.13	-.26**
$r_{AMIS, DDEV}$	-.76**	-.51**	-.81**	-.48**	-.35*	-.44*	-.53**
$r_{AMIS, QUAL}$	-.55**	-.41*	-.60**	.08	-.23	-.27	-.26**
$r_{RMIS, DISC}$	.24	-.10	-.34	-.12	.11	-.52*	-.13
$r_{RMIS, DDEV}$	-.34	.47**	-.20	.09	.13	.29	.02
$r_{RMIS, QUAL}$	-.44*	.42*	-.09	.15	.07	.41	.10
$r_{IMIS, DISC}$	-.07	.44*	.06	-.01	-.03	-.18	.10
$r_{IMIS, DDEV}$	.06	-.16	.09	.29	.17	.42	.07
$r_{IMIS, QUAL}$	.13	-.37*	.18	.20	.15	.37	.01

*Significant at the .10 level of probability
**Significant at the .05 level of probability